Evaluation of Clustering with Categorical and Mixed Type Variables and Cluster Number Determination

Löster, Tomáš
University of Economics, Prague, Department of Statistics and Probability,
W. Churchill Sq. 4
13067 Praha 3, Czech Republic
E-mail: tomas.loster@vse.cz

Řezanková, Hana
University of Economics, Prague, Department of Statistics and Probability,
W. Churchill Sq. 4
13067 Praha 3, Czech Republic
E-mail: hana.rezankova@vse.cz

1. INTRODUCTION

Cluster analysis involves a broad scale of techniques. Hence an important factor when examining data structure is therefore the comparison of resulting clusters obtained by various algorithms and selection of the best assignment of objects to clusters. Determining the optimal number of clusters is also important.

Current literature draws attention particularly to the evaluation of clustering in a situation when individual objects are characterized only by quantitative variables, see [2], [3]. The problems associated with the analysis of data characterized by qualitative or mixed type variables have only been dealt with to a limited extent. This is based on an analogy of the techniques applied when evaluating log-linear models for example.

In this paper we suggest new coefficients for the evaluation of resulting clusters based on the principle of the variability analysis. Furthermore, only coefficients for mixed type variables based on a combination of sample variance and one of the variability measures for nominal variables will be presented. Similar approaches can be applied in the case of qualitative variables while omitting the part characterizing the variability of quantitative variables.

The following text is organized in such a way that in Section 2 there is a description of the newly proposed coefficients and in Section 3 these coefficients are applied for determining the optimal number of clusters in real data files. Conclusion presents an evaluation of the obtained findings.

2. EVALUATION OF CLUSTERING RESULTS IN CASE OF MIXED TYPE VARIABLES

In this paper disjunctive clustering resulting in the unique assignment of objects to clusters is only considered. If objects are characterized only by qualitative variables it can be accomplished, for example, using hierarchical cluster analysis with the application of the coefficient of disagreement as a dissimilarity measure. In case of mixed type variables a log-likelihood distance measure can be applied (it is implemented in two-step cluster analysis in the IBM SPSS Statistics system).

The evaluation of the results of clustering can be based on within-cluster variability. The method is better which results in clusters with less variability. To determine variability in case that objects are characterized by mixed type variables, a combination of sample variance and *entropy* is applied in practice (in the SPSS system). Within-cluster variability for k clusters is determined by the formula

$$H(k)=\sum_{h=1}^{k}\frac{n_h}{n}\left(\sum_{t=1}^{m_1}\frac{1}{2}\ln(s_t^2+s_{ht}^2)+\sum_{t=1}^{m_2}\left(-\sum_{u=1}^{K_t}\left(\frac{n_{htu}}{n_h}\ln\frac{n_{htu}}{n_h}\right)\right)\right), \qquad (1)$$

where n is the number of objects, m_1 is the number of quantitative (continuous) variables, m_2 is the number of nominal variables, s_t^2 is the sample variance of the tth variable, s_{ht}^2 is the sample variance of the tth variable in the hth cluster, K_t is the number of categories of the tth variable, n_{htu} is the frequency of the uth category of the tth variable in the hth cluster, and n_h is the number of objects in the hth cluster.

We have proposed several coefficients for clustering evaluation both for the analysis with categorical variables (see [5]) and for mixed type variables (see [4]). In this paper we present the evaluation of some of them.

As an alternative to Formula (1) we suggest a measure which applies a combination of the sample variance and the *Gini coefficient*. For k clusters it can be determined according to the formula

$$G(k)=\sum_{h=1}^{k}\frac{n_h}{n}\left(\sum_{t=1}^{m_1}\frac{1}{2}\ln(s_t^2+s_{ht}^2)+\sum_{t=1}^{m_2}\left(1-\sum_{u=1}^{K_t}\left(\frac{n_{htu}}{n_h}\right)^2\right)\right). \qquad (2)$$

For determining the number of clusters we suggest to modify *the CHF index*. We can use either Formula (1) or Formula (2) as a variability measure, i.e. we obtain either *the CHFH index* in the form

$$I_{\mathrm{CHFH}}(k)=\frac{(n-k)\big(H(1)-H(k)\big)}{(k-1)H(k)}, \qquad (3)$$

or *the CHFG index* in the form

$$I_{\mathrm{CHFG}}(k)=\frac{(n-k)\big(G(1)-G(k)\big)}{(k-1)G(k)}. \qquad (4)$$

The high values of these indices indicate well separated clusters, i.e. the maximum value within a certain interval is searched.

The *Schwarz Bayesian information criterion* (BIC) can also be applied to determine the optimal number of clusters. It can be calculated according to the formula

$$I_{\mathrm{BIC}}(k)=2H(k)+k(2m_1+\sum_{t=1}^{m_2}(K_t-1)\ln(n). \qquad (5)$$

We newly suggest also used $G(k)$ instead of $H(k)$. This criterion will be denoted as I_{BICG} in the following text. The estimate of the number of clusters is determined on the basis of the minimum value of this coefficient.

3. APPLICATION OF NEW INDICES OF EVALUATION TO REAL DATA FILES

This part describes the results and conclusions of the practical application of the newly suggested coefficients applicable to mixed type variables. Two data files from *the UCI Machine Learning Repository* (http://archive.ics.uci.edu/ml/datasets.html) are analyzed. The BIC index is stated as a representant of the existing coefficients for a comparison with newly proposed indices.

3.1 The Wine File

The *Wine* file includes 178 wine samples. The original data file contains thirteen quantitative variables which express the the quantities of constituents. We created categories for two variables (*Flavanoids* and *Prolines*) in order to analyze the file with mixed type variables. For each of the wines the classification into some of three groups representing different cultivars is known.

Analysis of Clustering Results – Variant 1

In this part we present results of the analysis of the file with eleven quantitative variables and two recoded three-category variables. Table 1 shows the values of indices described in this paper. Three clusters which correspond to the correct number of groups were found as optimal on the basis of all indices.

Table 1: Evaluation of Variant 1

k	$I_{BIC}(k)$	$I_{CHFH}(k)$	$I_{BICG}(k)$	$I_{CHFG}(k)$
2	1896.95	55.88	1698.18	50.41
3	**1697.34**	**57.59**	**1555.43**	**52.19**
4	1730.45	46.37	1618.77	40.71
5	1788.90	39.90	1707.69	33.62
6	1867.52	35.24	1805.60	29.00
7	1945.47	32.47	1899.36	26.28
8	2035.82	30.07	1999.44	24.14
9	2134.30	28.02	2099.98	22.63
10	2239.16	26.21	2215.27	20.86
11	2350.72	24.52	2331.25	19.44
12	2458.29	23.36	2439.57	18.62
13	2567.92	22.36	2549.20	17.93
14	2678.14	21.54	2664.78	17.16
15	2798.98	20.44	2787.96	16.25

Table 2: Evaluation of Variant 2

k	$I_{BIC}(k)$	$I_{CHFH}(k)$	$I_{BICG}(k)$	$I_{CHFG}(k)$
2	2080.05	**52.90**	1834.91	39.01
3	1916.43	50.02	**1717.62**	**41.26**
4	**1912.55**	43.35	1743.15	36.13
5	1974.09	37.38	1842.76	29.80
6	2046.29	33.71	1929.39	26.91
7	2127.80	31.15	2033.13	24.36
8	2224.48	28.86	2142.56	22.39
9	2319.05	27.41	2259.34	20.68
10	2421.55	26.10	2373.33	19.52
11	2529.41	24.95	2490.09	18.55
12	2642.27	23.89	2615.63	17.47
13	2768.76	22.52	2740.57	16.62
14	2888.93	21.63	2864.52	15.96
15	3016.01	20.64	2995.29	15.19

Analysis of Clustering Results – Variant 2

In the second variant we analyzed the file with eleven quantitative variables and two recoded four-category variables. Table 2 shows the values of indices described in this paper. Four clusters were selected as optimal according to the known BIC criterion. On the basis of CHFH index two clusters were selected as optimal. It is therefore obvious that in these cases the correct number of clusters has not been determined. According to the BICG and CHFG indices (using a combination of sample variance and the Gini

coefficient and) three clusters were found as optimal, and thus the correct number was found.

When analysing this file, it was therefore found that the newly suggested indices based on a combination with the Gini coefficient can better determine the number of clusters than indices based on entropy.

3.2 The German Credit Data File

The *German Credit Data* file (the *Statlog* name is also cited) includes 1,000 objects (customers). The file contains seven quantitative variables (e.g. age in years, credit amount) and thirteen qualitative variables (e.g. personal status and sex, type of housing). For each of the customers the classification into some of two groups representing different level of risk is known.

Table 3: Evaluation of Credit Data Clustering

k	$I_{BIC}(k)$	$I_{CHFH}(k)$	$I_{BICG}(k)$	$I_{CHFG}(k)$
2	**22980.77**	**90.26**	**15418.42**	**75.85**
3	23085.31	69.37	15811.65	63.37
4	23357.68	60.45	16376.69	55.41
5	23669.31	56.17	17001.74	50.63
6	24137.86	52.05	17643.17	47.87
7	24739.15	48.05	18426.95	43.77
8	25371.55	45.03	19188.64	41.33
9	26059.56	42.37	20031.36	38.43
10	26800.67	39.91	20896.15	35.94
11	27561.56	37.85	21733.63	34.32
12	28372.80	35.82	22641.90	32.24
13	29160.13	34.33	23517.67	30.86
14	29930.50	33.22	24381.79	29.84
15	30769.17	31.86	25308.31	28.40

We analyzed the file with all variables. In Tables 3 there are values of all investigated indices. According to all indices two clusters were determined as optimal, which is the correct number.

4. CONCLUSION

In this paper we evaluated selected indices for determining the number of clusters when objects are characterized by mixed type variables. On the basis of real data files analyses we compared three newly proposed indices with the known BIC criterion. We knew the number of object groups and we were interested in agreement of the found optimal number of clusters with the

real number of groups. We had analyzed several (15) data files before and here we presented application of evaluation criteria to some them.

In two presented examples it was found that all indices determined the correct number of clusters, in one example only the criteria based on the Gini coefficient were successful. According to our experience when analyzing more data files, the CHFG index determines the correct number of clusters in most cases. The second successful criterion is the CHFH index. The BIC and BICG indices are less successful.

In our further research we will focus on some other evaluating criteria based on the principle of analysis of variance and the R-squared coefficient. Further, we will be interested in evaluation of different methods for clustering objects characterized by qualitative or mixed type variables, including latent class models, for example.

Acknowledgements. This work was supported by projects GACR P202/10/0262, MSM6138439910, and IGA VSE F4/5/2011.

REFERENCES

[1] Calinski, T., Harabasz, J.: A dendrite method for cluster analysis, *Comunications in Statistics*, Vol. 3, 1974, 1–27.

[2] Gan, G., Ma, C., Wu, J.: *Data Clustering Theory, Algorithms, and Applications*. ASA, Philadelphia, 2007.

[3] Halkidi, M., Batistakis, Y., Vazirgiannis, M.: *Clustering Algorithms and Validity Measures.* SSDBM, Athens, 2001.

[4] Řezanková, H., Húsek, D., Löster, T.: Clustering with mixed type variables and determination of cluster numbers. In: *COMPSTAT 2010*.CNAM and INRIA, Paris, 2010, 1525–1532.

[5] Řezanková, H., Löster, T., Húsek, D.: Evaluation of categorical data clustering. In: *Advances in Intelligent Web Mastering – 3*. Springer Verlag, Berlin, 2011, 173–182.